EXHIBIT A

Efficient realization of iSCSI Data Acknowledgment on RDMA fabrics

Mallikarjun Chadalapaka
Phone: +1 916-785-5621
Email: cbm@rose.hp.com

Networked Storage Architecture
Network Storage Solutions
Hewlett-Packard Company
Roseville CA USA

Abstract:

This paper discusses a new technique for efficient realization of iSCSI data acknowledgement requests (A-bit on Data-in PDUs) when iSCSI is operating on top of a Datamover protocol such as iSER. The traditional iSCSI data acknowledgment model involves the iSCSI layer on the target seeking a data acknowledgment from the initiator via the aforementioned A-bit, and the iSCSI layer on the initiator responding to the same via an iSCSI SNACK PDU. This traditional model, if adopted without changes to the RDMA Data movers, will lead to an extremely inefficient realization of the data acknowledgment feature in iSCSI/RDMA-Datamover implementations such as iSCSI/iSER. This paper describes a different technique that takes advantage of RDMA ordering rules and iSCSI error recovery techniques in realizing the same iSCSI data acknowledgment model in a highly optimized way.

1. The iSCSI data acknowledgment model

The iSCSI protocol specification defines a data acknowledgment model that is primarily intended for iSCSI targets to efficiently manage their buffers in responding to SCSI Read commands. The idea is that by proactively seeking a data acknowledgment from the iSCSI layer on the initiator and receiving it, the iSCSI layer on the target can be certain that the just acknowledged data will not be requested for a retransmission down the road by the iSCSI layer on the initiator. This assurance could greatly aid the iSCSI and SCSI layers on the target because the data buffers don't have to be kept around until the end of the task, but can be immediately reused for other I/O demands.

The iSCSI layer on the target conveys the data acknowledgment request to the initiator iSCSI layer via the setting of the A-bit in the iSCSI Data-in PDU from the target. Upon receiving a valid data acknowledgment request via the A-bit, the initiator iSCSI layer in turn is required to respond back with a data acknowledgment response in the form of an iSCSI SNACK PDU. This request-response interaction thus will acknowledge all the data up to and including the Data-in PDU that had the A-bit set.

Note that the whole iSCSI data acknowledgment model is usable only if the operational ErrorRecoveryLevel of the iSCSI session is greater than 0. An ErrorRecoveryLevel of 0 means that there's no data recovery, so a data acknowledgment is not useful by definition. An ErrorRecoveryLevel of 1 means that data recovery is supported on that iSCSI session, so the iSCSI layer on the target could deploy data acknowledgments to realize some efficiency in local buffer management. An ErrorRecoveryLevel of 2 implies task failover is also supported in addition to all that of ErrorRecoveryLevel=1, so the iSCSI layer on the target may still use data acknowledgments for the same reason.

2. The RDMA Datamover model

In the RDMA Datamover model (such as the one employed by the iSER protocol running on the iWARP protocol suite), the Datamover layers on the initiator and the target are responsible for moving all the data unbeknownst to the iSCSI layer on the initiator. In the case of a SCSI Read command, the iSER layer on the target simply moves all the Read data using one or more RDMA Writes back to the initiator RNIC. The key benefit of this data movement model is that the iSCSI layer on the initiator receives only a single interrupt at the conclusion of any SCSI Command (a Read command in this scenario), because it is not involved in the data movement.

3. Data acknowledgment problems with the RDMA model

As much as the RDMA Datamover model is compelling, it poses certain problems for the data acknowledgment expectations of the iSCSI specification. The biggest problem is that the iSCSI layer on the initiator cannot acknowledge any data back to the target iSCSI layer since the initiator iSCSI layer was not involved in the data movement to begin with! The second biggest problem is that the iSCSI layer on the initiator needs to be interrupted in order to respond to the acknowledgment request, and this violates the single interrupt per I/O model. This multiple interrupt problem is further compounded for the initiators because there is no limit on the number of acknowledgments that may be sought during the course of a SCSI Read command. Thus the number of interrupts an initiator may have to field for a long-running I/O is unbounded.

4. Proposed solution

If the data acknowledgment request originated by the iSCSI target layer can be couched in a generic form (i.e. without iSCSI or iSER specifics) so that it may be automatically responded to by the initiator's RDMA-capable Network Interface Controller (RNIC), it solves the problems mentioned in 3.

An RDMA Read Request is a good "generic form" of an acknowledgement request. An RDMA Read Request-Read Response pair is analogous in an abstract sense to a data acknowledgment request-response pair, and the RNIC receiving the RDMA Read Request on the *initiator* can automatically respond (without generating a local interrupt to the iSER layer) with an RDMA Read Response. Hence the proposal is to use an RDMA Read Request to "read" zero bytes out of the initiator's memory, whenever the A-bit is set on a Data-in PDU coming down from the iSCSI layer on the target. The RDMA Write/Read ordering rules of the RDMA Protocol ensure that the RDMA Read Request will not pass the RDMA Write Request and so the RDMA Read Request essentially acts to "flush" the connection of all the preceding RDMA Writes carrying the SCSI Data-in. The iSER layer on the target, when it receives the RDMA Read Response, can generate a notification back to the local iSCSI layer notifying it of the arrival of the data acknowledgment, essentially mimicking the SNACK-based data acknowledgment response.

It turns out that there is a further optimization possible here based on the operational ErrorRecoveryLevel of the iSCSI session. This optimization will simplify the wire protocol and iSER layer-to-RNIC interactions on the *target* in realizing the iSCSI data acknowledgment. If the operational ErrorRecoveryLevel is 1, as described in section 1, the connection recovery feature of the iSCSI protocol cannot be used. This further implies that if the connection fails for any reason, the data associated with the tasks on the failed connection will not be requested on a new connection since the tasks cannot be failed over. The RNIC interface guarantees that once the completion message for an RDMA operation is delivered to the iSER layer, the local data buffers associated with that RDMA operation will not be accessed by the RNIC and the associated data will be transferred if the connection stays up,. The combined implication is that when the operational ErrorRecoveryLevel=1, the iSER layer on the target can simply mimic a data acknowledgment response (as if received from the initiator) based on the RNIC-local completion message of the RDMA Write operation associated with the SCSI Data-in.

The pseudo-code of the proposed algorithm thus would look as describe din section 5.

5. Algorithm for the target iSER layer

```
If (the A-bit is set on the SCSI Data-in PDU) then
        If (the operational ErrorRecoveryLevel=2 or if ErrorRecoveryLevel is unknown) then
                Generate the standard RDMA Write for the SCSI Data-in PDU.
                Generate a zero-length RDMA Read Request after the RDMA Write.
                 Wait for the RDMA Read Response arrival
        else if (the operational ErrorRecoveryLevel=1) then
                Generate the standard RDMA Write for the SCSI Data-in PDU.
                Wait for the local RDMA Write Completion
        endif
endif
```

Once the event being waited for – RDMA Read Response arrival or the local RDMA Write completion – occurs, the iSER layer on the target must generate a data acknowledgment notification to the iSCSI layer. This completes the iSCSI data acknowledgment expectations as far as the target iSCSI layer is concerned. Note that the initiator iSER or iSCSI layers need no special handling or logic in this proposed model.

6. Conclusion

This paper discusses the problems faced in meeting the iSCSI data acknowledgment expectations in the context of an RDMA Datamover and proposes a solution to the problem. The proposed solution preserves the "single interrupt for SCSI command" model on the initiator, even while meeting the data acknowledgment needs of the iSCSI layer on a target. The proposed solution in addition also includes a performance optimization on the target side that will cut down the wire protocol exchanges and speeds up the data acknowledgment response back to the iSCSI layer.